Supplementary Materials

Exome Sequencing-Based Copy-Number Variation and Loss of Heterozygosity Detection: ExomeCNV

J. Fah Sathirapongsasuti[1,2,3]*, Hane Lee*[3,4]*, Basil A.J. Horst*[4,5,6]*, Georg Brunner*[7]*, Alistair J. Cochran*[4]*, Scott Binder*[4]*, John Quackenbush*[1,2]*, Stanley F. Nelson*[3,4,*]

[1] *Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115*

[2] *Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115*

[3] *Department of Human Genetics and* [4] *Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA 90095*

[5] *Department of Dermatology and* [6] *Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY 10032*

[7] *Department of Cancer Research, Skin Cancer Center Hornheide, Münster, Germany*

[*] *To whom correspondence should be addressed.*

Biases in Exon Capture Process

We start with an observation that if the exome capture has no bias and the distribution of reads is uniform across the exome, the distribution of depth of coverage (average number of reads per basepair) will appear to follow a Poisson distribution, with mean equal to variance. However, if exon capture biases are present, the variance to mean ratio will inflate above 1, a situation called overdispersion. Thus, we use the variance to mean ratio, called overdispersion factor ϕ, as a measure of the exon capture bias. The effect of known sources of bias, i.e. GC-content and sequence mapability, is considered and a final justification for using paried-sample comparison approach is given. All of the data presented here are based on the melanoma samples described in the paper and the methods.

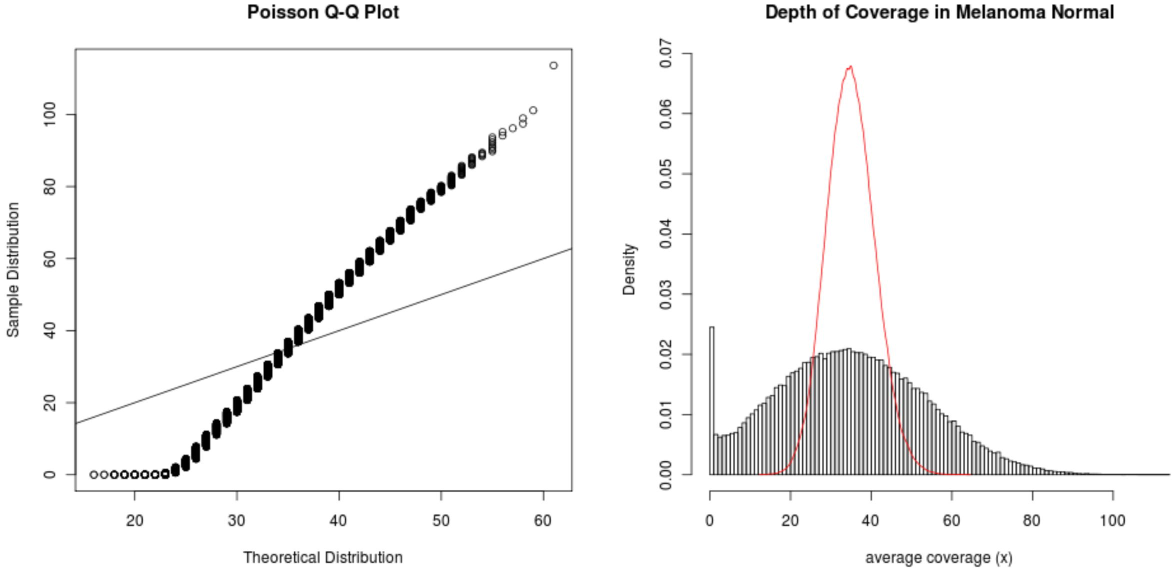


Figure S1 Poisson Q-Q plot and histogram shows poor fit to Poisson distribution and a high level of overdispersion.

Within one sample, there is a substantial amount of biases, and this is reflected through the Poisson Q-Q plot (Figure S1) and the overdispersion factor of 9.5. Part of this variability comes from the exons with zero coverage. We have examined theses exons with no coverage (available at http://genome.ucla.edu/~fah/ExomeCNV/supplement/SureSelect_No_Coverage_Exons_G3362.bed) and found that they correspond to homologs or repetitive sequences and have low mapability scores (based on ENCODE CRG mapability score). The aligner program discarded reads mapped to these regions because of the ambiguity. Even when removing these zero coverage exons, the amount of bias observed remains substantial (overdispersion factor of 8.9).

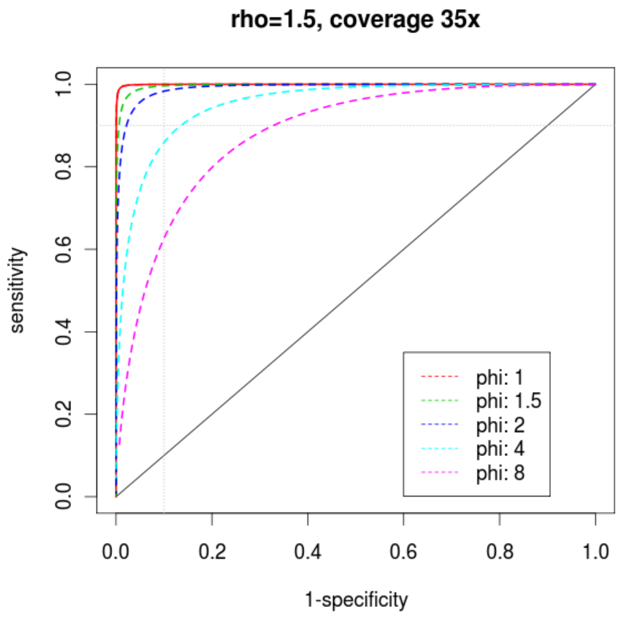


Figure S2 Effect of overdispersion on specificity and sensitivity of detecting a duplication event on an exon of size 500bp with 35x depth-of-coverage. The solid red curve is the ROC curve for the case where there is no overdispersion, and the subsequent curves are for the cases with increasing overdispersion factors.

Effect of Overdispersion on Specificity and Sensitivity of ExomeCNV

The effect of overdispersion to Poisson distribution can be modeled through the quasi-likelihood approach in which the variance is allowed to inflate: $\sigma^2 = \phi\ \mu$. Thus the transformed depth-of-coverage ratio statistic (see Methods for definition and derivation) becomes:

$$t(\rho,\phi) = \frac{\mu_Y R - \mu_X}{\sqrt{\sigma_Y^2 R^2 + \sigma_X^2}} = \frac{\lambda R - \rho\lambda}{\sqrt{\phi\lambda R^2 + \rho\phi\lambda}} = \frac{(R-\rho)}{\sqrt{R^2+\rho}}\sqrt{\frac{\lambda}{\phi}}.$$

That is, the overdispersion factor ϕ affects the statistic t by directly scaling down the average depth-of-coverage λ by a factor of ϕ. In other words, overdispersion will reduce the power and accuracy of prediction as though the depth-of-coverage is reduced by the factor ϕ. For example, an exon with 35x depth-of-coverage and an overdispersion factor of 5 will have the same power of prediction as an exon with 7x coverage but no overdispersion. Figure S2 illustrates the effect on the ROC curves for prediction of a duplication event on an exon size 500bp with 35x depth-of-coverage.

However, it is noteworthy that this calculation is only true under an assumption that taking the ratio of depth-of-coverage does not get rid of the overdispersion (the biases). We will soon see that this is fortunately not the case, and taking the ratio of depth-of-coverage of the same exon does significantly reduce the overdispersion effect.

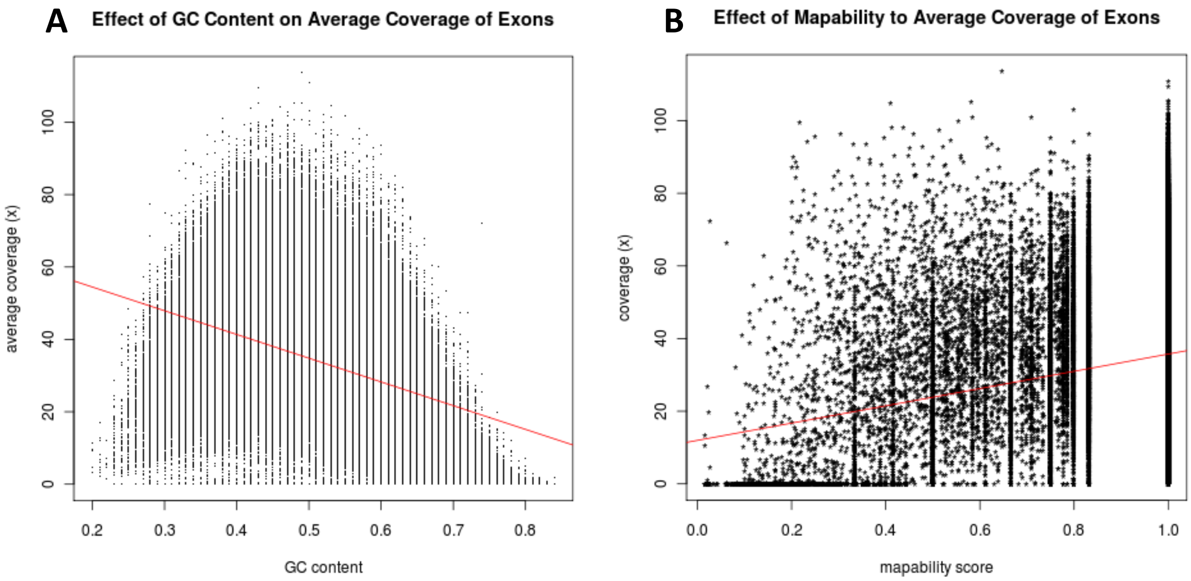


Figure S3 Effect of GC-content and mapability on depth-of-coverage of exons.

Effect of GC-Content

Figure S3A shows that sequences with extreme GC contents (especially high GC-content) tend to have lower depth-of-coverage. This is in agreement with previous sequencing projects [1-2]. By linear regression, GC-content accounts for 38.52% of the variability in the depth-of-coverage. Correcting for GC-content, we managed to reduce the overdispersion factor to 6.

Effect of Mapability

Mapability, or uniqueness of DNA sequence, can affect the efficiency of sequence alignment. Repetitive sequences have low mapability score and tend to be harder to align, resulting in generally lower depth-of-coverage. On the other hand, high mapability scores indicate unique sequences, and higher depth-of-coverage is observed. Mapability scores (CRG GEM-Alignability of 36mers with no more than 2 mismatches) were retrieved from UCSC Genome Browser, and an average score is calculated for each exon. Figure S3B shows the effect of mapability score on depth-of-coverage, with higher scores associated with higher coverage. Ordinary linear regression suggests that mapability accounts for 2.43% of the variance in depth-of-coverage distribution. Correcting for both GC-content and mapability reduces the overdispersion factor to 4.

Taking the Ratio of Depth-of-Coverage Reduces Probe-specific Biases

Although GC-content and mapability can help explain some of the extra variability in depth-of-coverage, more biases remain unexplained. Since our exon capture was done using microarray probes, probe-specific biases such as capture efficiency can have strong effect on the number of mapped reads. Because of the lack of mean to measure and correct for these biases directly, we considered the usefulness of the indirect approach of taking the ratio of depth-of-coverage. Our assumption is that if the effect of biases on depth-of-coverage of an exon is consistent across samples, taking the ratio of depth-of-coverage of the exon from two independent samples will reduce the biases. We looked at the distribution of depth-of-coverage of an exon across 10 exomes available internally and measured the overdispersion factor of the exon. Each overdispersion factor measures how well the depth-of-coverage of the exon follows Poisson distribution, and it turned out that most of the exons have overdispersion factors of less than 1 (mean 0.91, median 0.74, and 3rd quartile 1.17). This implies that it is reasonable to model depth-of-coverage by Poisson distribution and that taking the depth-of-coverage ratio from two samples will reduce the biases.

ROC of detecting deletion and duplication

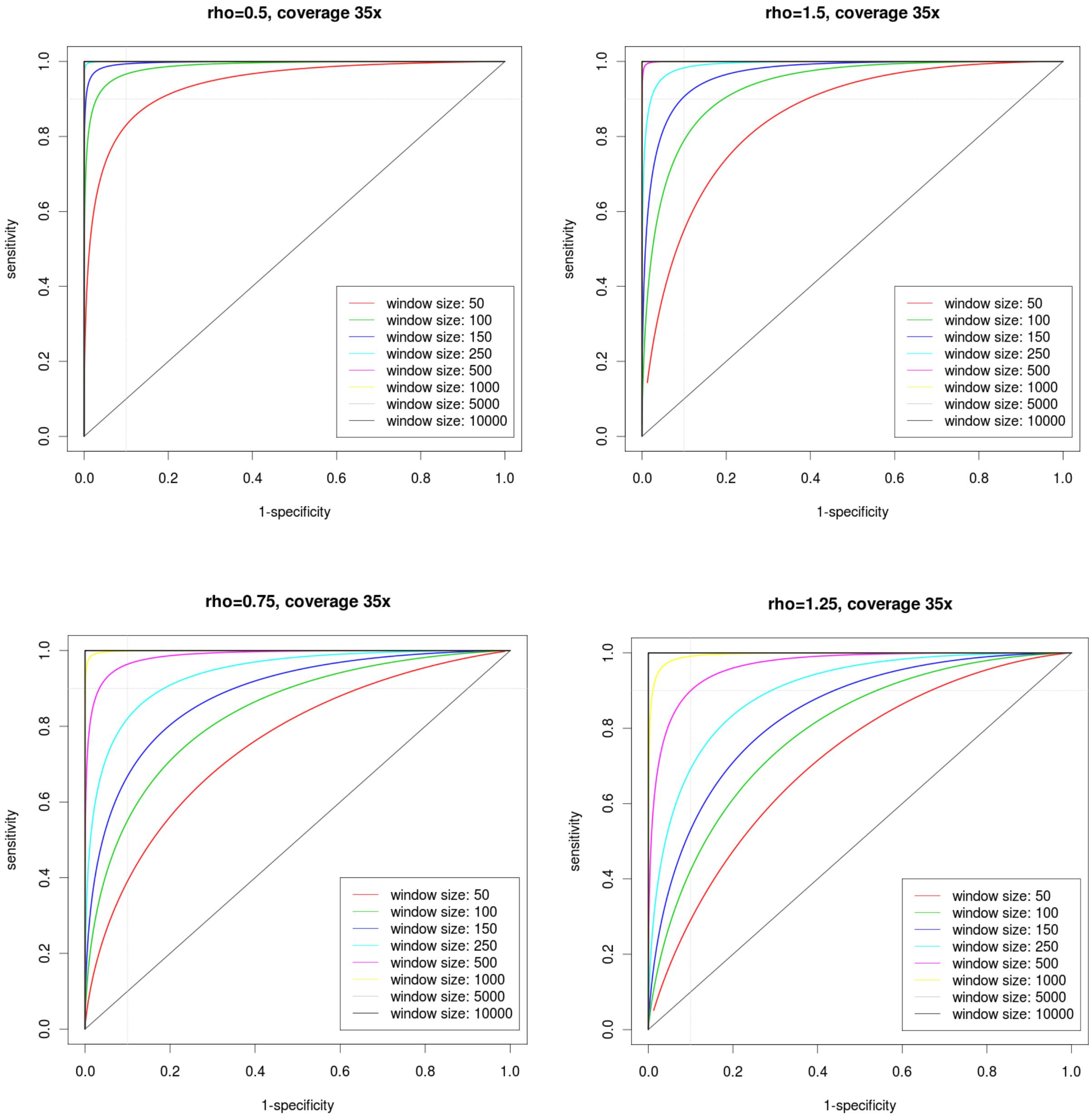


Figure S4 ROC curves for detecting (a,c) deletion and (b,d) one copy duplication. The depth-of-coverage is fixed at 35x and read length at 70bp. The dotted gray lines correspond to 95% specificity and sensitivity. Generally, it is more difficult to detect amplification than it is deletion. And at 35x coverage, deletion event (a) can be detected at 95% sensitivity and

specificity for exons of size 100 or less while duplication (b) can be detected for exons of size 250 or less. In the presence of admixture of normal genomes, the copy-number ratio ρ tends to one. For example, at the admixture rate of 50%, ρ = 0.75 for deletion and ρ = 1.25 for duplication. The power and sensitivity decreases as a result of the admixture, and the ROC curves show that (c) deletion can be detected at 95% sensitivity and specificity for exons of size 500bp or less and (d) duplication can be detected for exons of size less than 500bp.

Justification for the depth-of-coverage ratio threshold used for calling CNV

In calling CNV, we need to identify a cutoff $r(\rho)$ which yields desired minimum specificity and/or sensitivity for testing a particular copy-number ratio ρ at a particular exon with some depth-of-coverage and length. Solving the equations in the main paper Section 2.1 for R, we derive the cutoff value for an α-level test:

$$r_\alpha(1) = \frac{\lambda + t_\alpha\sqrt{2\lambda - t_\alpha^2}}{\lambda - t_\alpha^2},$$

where

$$t_\alpha = \begin{cases} \Phi^{-1}(\alpha) & \text{if } \rho < 1, \\ \Phi^{-1}(1-\alpha) & \text{if } \rho \geq 1 \end{cases}$$

And the cutoff value for a test of power at least 1 – β is:

$$r_\beta(\rho) = \frac{\rho\lambda + t_\beta\sqrt{\rho(\lambda - t_\beta^2 + \rho\lambda)}}{\lambda - t_\beta^2},$$

where

$$t_\beta = \begin{cases} \Phi^{-1}(1-\beta) & \text{if } \rho < 1, \\ \Phi^{-1}(\beta) & \text{if } \rho \geq 1 \end{cases}$$

If $\rho > 1$, that is we are considering a duplication event, a test rejects when the observed coverage ratio $R > r_{\text{cutoff}}$. Conversely if $\rho < 1$, that is we are considering a deletion event, a test rejects when R < r_{cutoff}.

Figure S5 shows a graphical representation of the relationship among α, β, $r_\alpha(1)$, and $r_\beta(\rho)$. From Figure S5, we note that the exon represented by the red curve does not have sufficient coverage to achieve the desired specificity (1 – α) and sensitivity (1 – β) simultaneously, whereas the exon represented by the green curve does have sufficient coverage. If $\rho > 1$, R increases in the direction of the arrows, and the inequality $r_\alpha(1) > r_\beta(\rho)$ indicates sufficiency of the coverage (as is the case for the green line), and the reverse indicates insufficiency of the coverage (as is the case for the red line). If an exon does not have sufficient coverage, we refrain from declaring CNV for that exon.

In the case when an exon has enough coverage to call CNV, there are multiple possible cutoff values r ε $[r_\beta(\rho), r_\alpha(1)]$, and one can choose to optimize r for sensitivity, specificity, or a function of the two, e.g. area under curve (AUC) = (sensitivity + specificity)/2. We allow the user to choose which option to optimize when performing the test as each option is suitable for different applications.

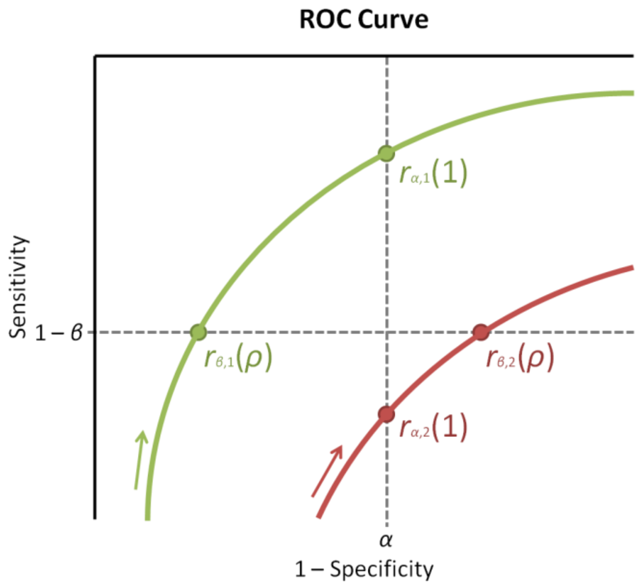


Figure S5 A schematic sketch showing relationship among α, β, $r_\alpha(1)$, and $r_\beta(\rho)$. The red curve is an ROC curve of an exon with insufficient coverage to achieve desired specificity and sensitivity $1 - \alpha$ and $1 - \beta$, while the green curve is one with sufficient coverage. The arrows indicate the direction in which the copy-number ratio R increases. Thus, an exon has sufficient coverage to call CNV when $r_\alpha(1) > r_\beta(\rho)$, and not otherwise.

Sequential Merging Procedure

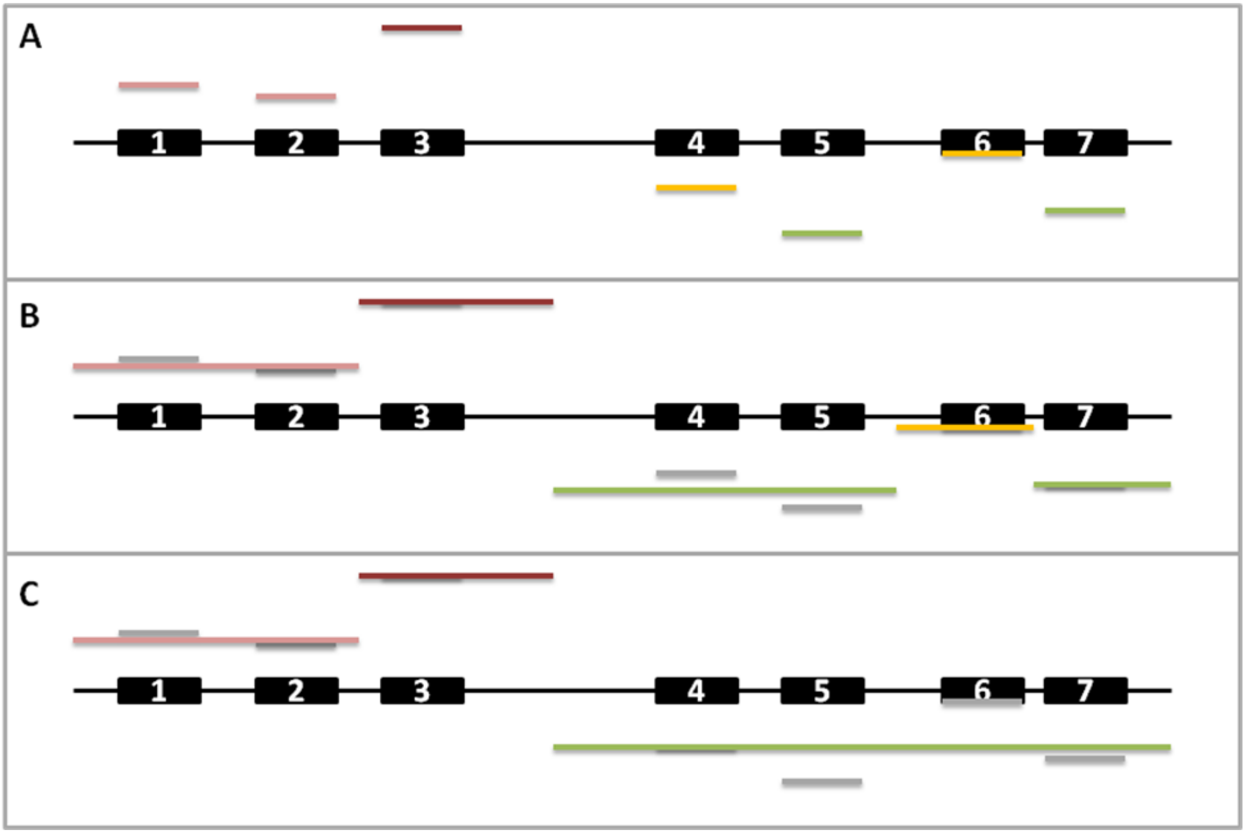


Figure S6 Sequential Merging Procedure.

This illustrates three levels of granularity of segmentation. (A) At the finest level of granularity, copy-number variation is assessed at each exon individually. Exons 1 and 2 are called as duplicated (copy number=3) while Exon 3 is called as amplified (copy number > 3). Exons 5 and 7 are called as deleted while Exons 4 and 6 are called as copy-neutral, either because they are truly copy-neutral or because of the lack of power. (B) Circular Binary Segmentation (CBS) subdivides the genome into segments at fine granularity. Exons 1 and 2 merge into one duplicated segment while Exon 4 and 5 merge into a deleted segment. With the merged segment at this level, there is enough power to call deletion on Exon 4. (C) Final segmentation at the coarsest level. Exons 4-7 merge together into a large deleted segment, giving power to call deletion. The final CNV calls consist of three segments as shown.

Other Statistics for Detecting Segmental LOH

Here we discuss the choice of statistics used to detect LOH in a segment produced by the circular binary segmentation (CBS) algorithm. An F-test for equality of variance described in the Methods Section appears to be very sensitive and can detect slight changes in copy-number, which are sometimes caused by non-LOH events such as amplification. Since the increase in variance due to LOH is generally greater than those due to non-LOH events, we can compensate for this over-sensitivity with a conservative p-value threshold. However, since the choice of p-value threshold is quite arbitrary, we considered other statistics: non-parametric Wilcoxon rank-sum test and folded-normal test.

One major challenge in detecting LOH is the fact that non-reference or B-alleles are not always on the same chromosome strand. Thus, the direction of the deviation of B-allele frequencies (BAFs) from 0.5 cannot be used, and the BAFs cannot be combined directly in any meaningful way. The only information we can gain from BAFs is the magnitude of the deviation: $|BAF - 0.5|$ or $|BAF_{\text{case}} - BAF_{\text{control}}|$ (Figure S8). For a non-LOH region, we may assume that BAF is normally distributed with mean 0.5 and certain variance σ^2 (Figure S7A), and so $|BAF - 0.5|$ and $|BAF_{\text{case}} - BAF_{\text{control}}|$ follows the corresponding folded-normal distribution (Figure S7C,D). For an LOH region, BAF will have a bimodal distribution centering around 0.5, with each half approximable by a normal distribution (Figure S7B). These observations serve as basis on which Wilcoxon test and folded-normal test are developed.

The motivation for using Wilcoxon test arises from an observation that the deviations of BAF $|BAF - 0.5|$ in LOH and non-LOH regions follow two distributions with different means (Figure S8B). Thus we can use Wilcoxon rank-sum test, which is non-parametric and insensitive to model assumption, to detect the difference in the means of the BAF deviation between case and control. The Wilcoxon test appears to be very sensitive, achieving as high as 96.89% sensitivity in detecting LOH in the melanoma sample. However, because of its sensitivity, it also detects changes in the deviation of BAF due to other non-LOH copy-number changes, resulting in low specificity (50-61%). In general, the Wilcoxon test performs very similarly to the F-test.

We attempt to address over-sensitivity problem encountered by the Wilcoxon test and F-test by developing a test that is less sensitive and is able to distinguish between LOH and non-LOH BAF shifts. Because the absolute difference in BAF $|BAF_{\text{case}} - BAF_{\text{control}}|$ can be assumed to follow a folded-normal distribution, we can use the folded-normal distribution to test for a significant deviation from the null case where BAF_{case} and BAF_{control} are identically distributed. The mathematical details of the test are outlined here:

Assuming BAF_{case} and BAF_{control} follow the normal distributions $N(\mu_{\text{case}}, \sigma^2)$ and $N(\mu_{\text{control}}, \sigma^2)$, respectively, the difference $BAF_{\text{case}} - BAF_{\text{control}}$ follows $N(\mu_{\text{case}} - \mu_{\text{control}}, 2\sigma^2)$, and the absolute difference $|BAF_{\text{case}} - BAF_{\text{control}}|$ follows a folded-normal distribution with mean:

$$E(|BAF_{\text{case}} - BAF_{\text{control}}|) = \sigma\sqrt{2/\pi}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left[1 - 2\Phi\left(\frac{\mu}{\sigma}\right)\right]$$

and variance:

$$\mathrm{Var}(|BAF_{\mathrm{case}} - BAF_{\mathrm{control}}|) = \mu^2 + \sigma^2 + \left\{\sigma\sqrt{2/\pi}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left[1 - 2\Phi\left(\frac{\mu}{\sigma}\right)\right]\right\}^2$$

where $\mu = \mu_{\mathrm{case}} - \mu_{\mathrm{control}}$, and Φ is the standard normal cumulative distribution function (CDF).

When BAF_{case} and BAF_{control} are identically distributed, $\mu = \mu_{\mathrm{case}} - \mu_{\mathrm{control}} = 0$, i.e. a half-normal distribution. Thus, the CDF of the folded-normal becomes:

$$P(X \leq x) = 1 - P(X > x) = 1 - 2P(Z > x) = 2\big(1 - P(Z > x)\big) - 1 = 2\Phi(x) - 1$$

and the p-value for the folded-normal test is given:

$$P(X > x) = 1 - P(X \leq x) = 2\big(1 - \Phi(x)\big)$$

where x is a realization of the standardized $X = |BAF_{\mathrm{case}} - BAF_{\mathrm{control}}|/\sigma$. In practice, we use $x = \mathrm{average}(|BAF_{case} - BAF_{control}|)/\mathrm{s.e.}(|BAF_{\mathrm{case}} - BAF_{\mathrm{control}}|)$.

As expected the folded-normal test gives more conservative results, achieving 97.52% specificity with 54.27% sensitivity. When combined with the position-wise binomial test (described in the main text), we can improve the sensitivity to 67.55% while lowering specificity to 88.01%.

The choice of test for LOH depends on the application and users' tolerance to false positive and false negative. Finally, we believe that there exists a more efficient test that makes use of other information, such as predicted CNV status and haplotype, and more sophisticated computational techniques, such as Hidden-Markov Model (HMM), but we consider this to be beyond the scope of our present study.

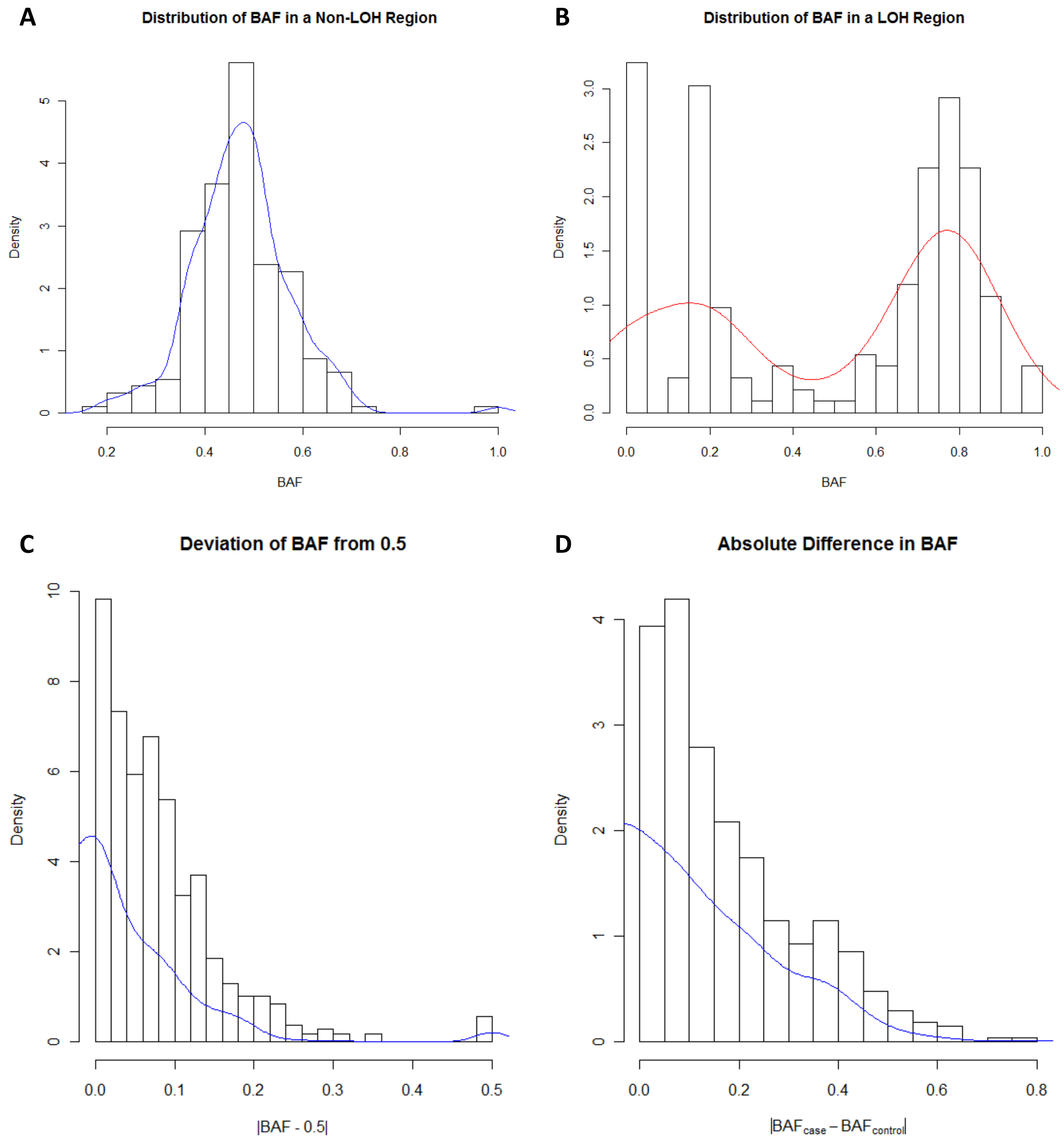


Figure S7 Distribution of B-allele Frequencies (BAF) in non-LOH and LOH regions. (A) The distribution of BAF in a non-LOH region follows a normal distribution with mean 0.5, while (B) the distribution of BAF in an LOH region follows a bimodal mixture of normal. (C), (D) The deviation of BAF from 0.5 $|BAF - 0.5|$ and the absolute difference in BAF ($|BAF_{case} - BAF_{control}|$) follow folded-normal distributions.

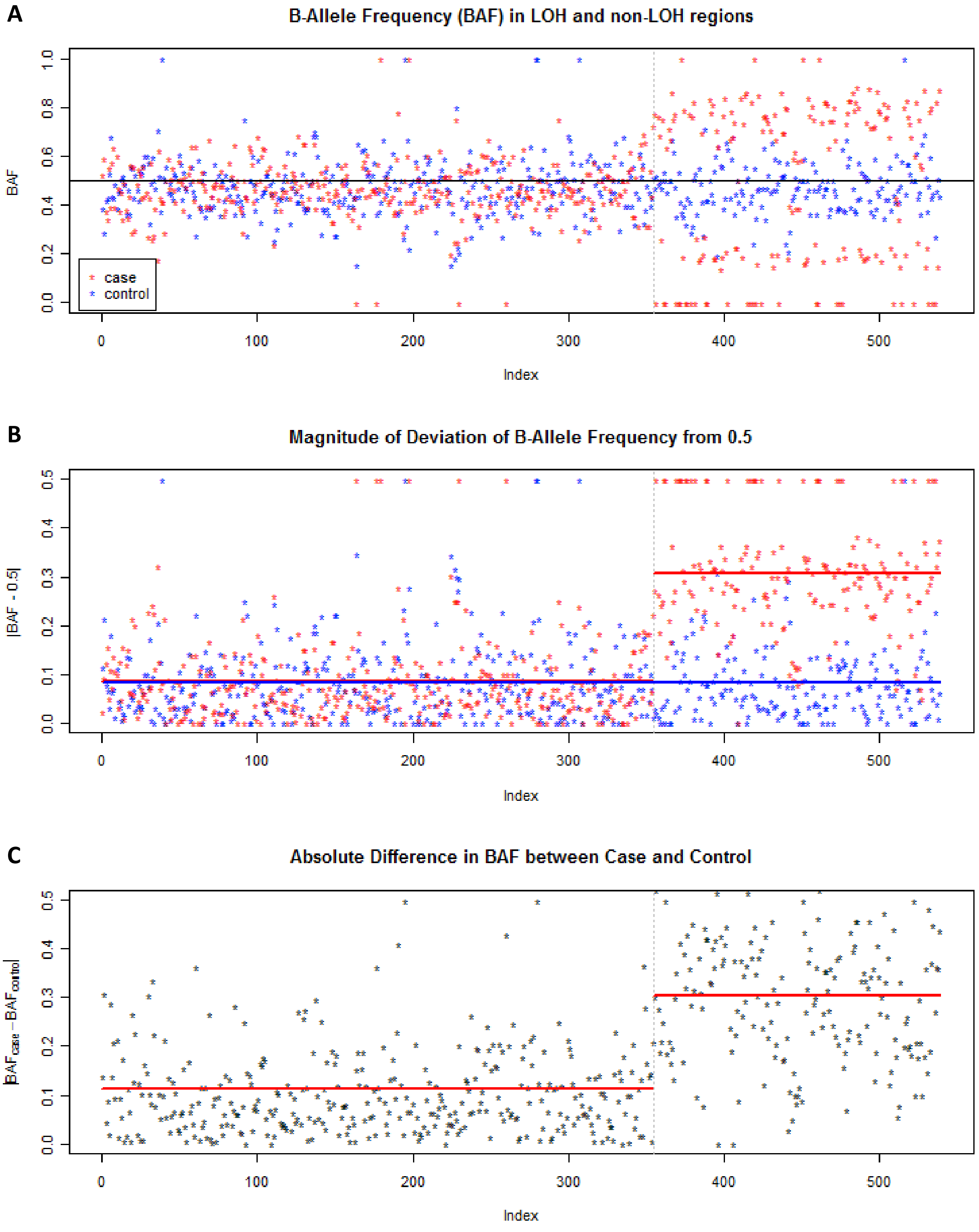


Figure S8 BAF, Deviation of BAF from 0.5, and Absolute Difference in BAF between Case and Control.

Size Distribution of CNV Segments

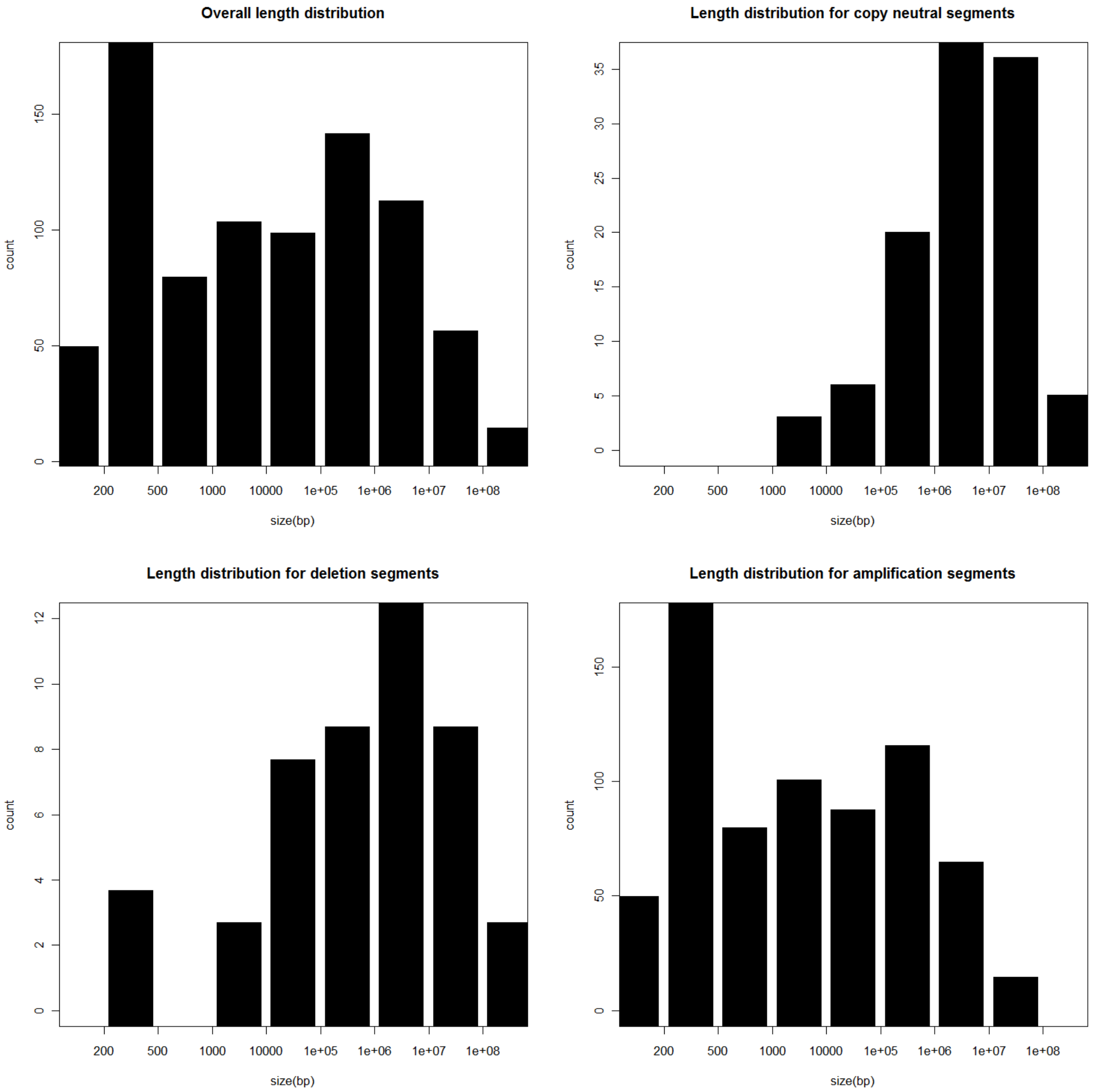


Figure S9 Size distribution of CNV calls in the melanoma samples. The sizes of the CNV segments from ExomeCNV range from single exon (120bp) to whole chromosome (chr 10 and 18). The distribution of amplified segments are biased toward smaller segments because ExomeCNV distinguishes segments with evidence for higher copy numbers (e.g. 3, 4, 5) apart while considering all deleted segments the same. Thus all the deleted segments were merged together forming larger segments while amplified segments remain fragmented. We cannot verify the biological validity of this behavior at this point.

In the analysis of melanoma samples, ExomeCNV was allowed to call higher copy numbers (3, 4, 5, etc.). These higher number amplifications were observed in small segments, often single exon, and were not merged together during the CBS-sequential merging step. Thus, the amplified segments have higher number of small segments (< 500bp). In our validation, we treated these higher copy number amplifications as one group.

Lane1-Lane2 Test for False Positive

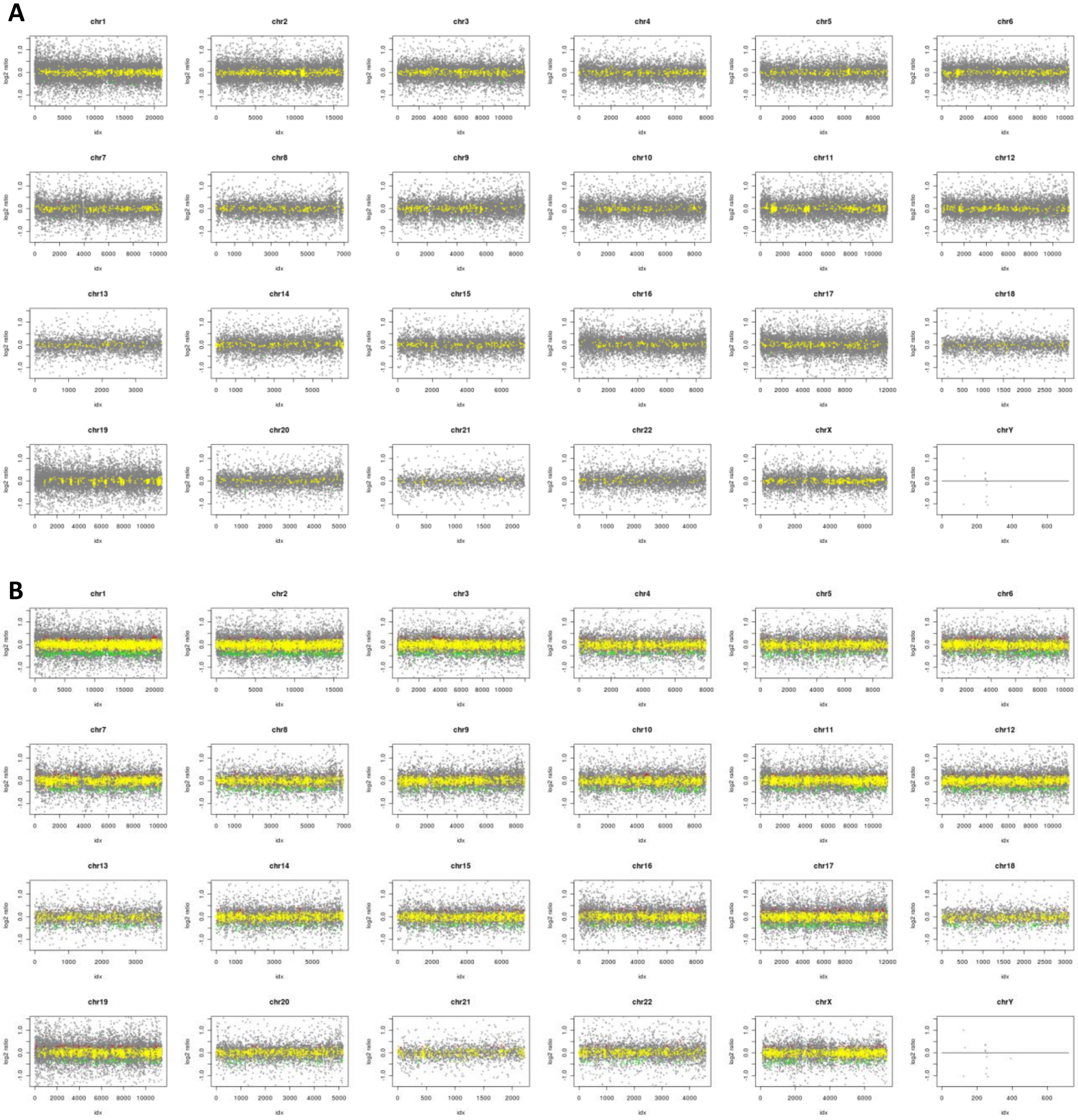


Figure S10 Lane1-Lane2 Test for False Positive. The plots show log depth-of-coverage ratio and CNV calls (yellow = copy-neutral, red = amplification, green = deletion) by chromosome. Gray dots are the exons with insufficient coverage to be called on their own. (A) is the results from setting minimum specificity to 0.99 and (B) from setting minimum specificity to 0.9. Note that (A) yields higher sensitivity but also lowers the sensitivity by calling CNV for a less number of exons.

We processed sequencing data from two lanes of the same matched skin sample run, call these Lane1 and Lane2. The average depth-of-coverage of Lane1 and Lane2 are comparable (21.3x and 20.4x, respectively). Since Lane1 and Lane2 were from the same library and processed exactly the same way, Lane1 should have no copy-number change with respect to Lane2, and any CNV calls are false positives. We ran ExomeCNV, treating Lane1 as case and Lane2 as control, and counted the number of exons falsely classified as deleted or amplified. Here we set estimated admixture rate to 30%. Adjusting minimum specificity from 0.9 to 0.999, we observe empirical specificities as reported in the main text (Figure S10). When the minimum specificity was set to 0.999, ExomeCNV made 2865 calls; all of which were copy neutral. At minimum specificity of 0.99, ExomeCNV made 5738 calls: 25 deletions, 5711 copy-neutral, and two duplications. And at minimum specificity of 0.9, 27366 calls were made: 1903 deletions, 25073 copy-neutral, and 390 duplications.

Sex-Chromosome Test for False Negative

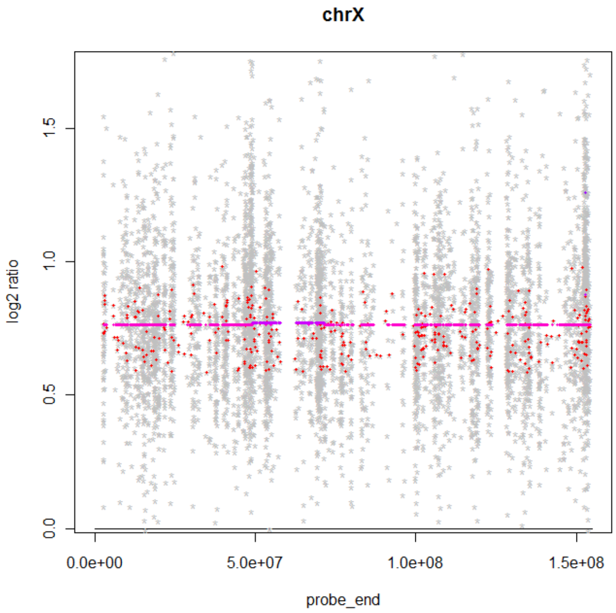
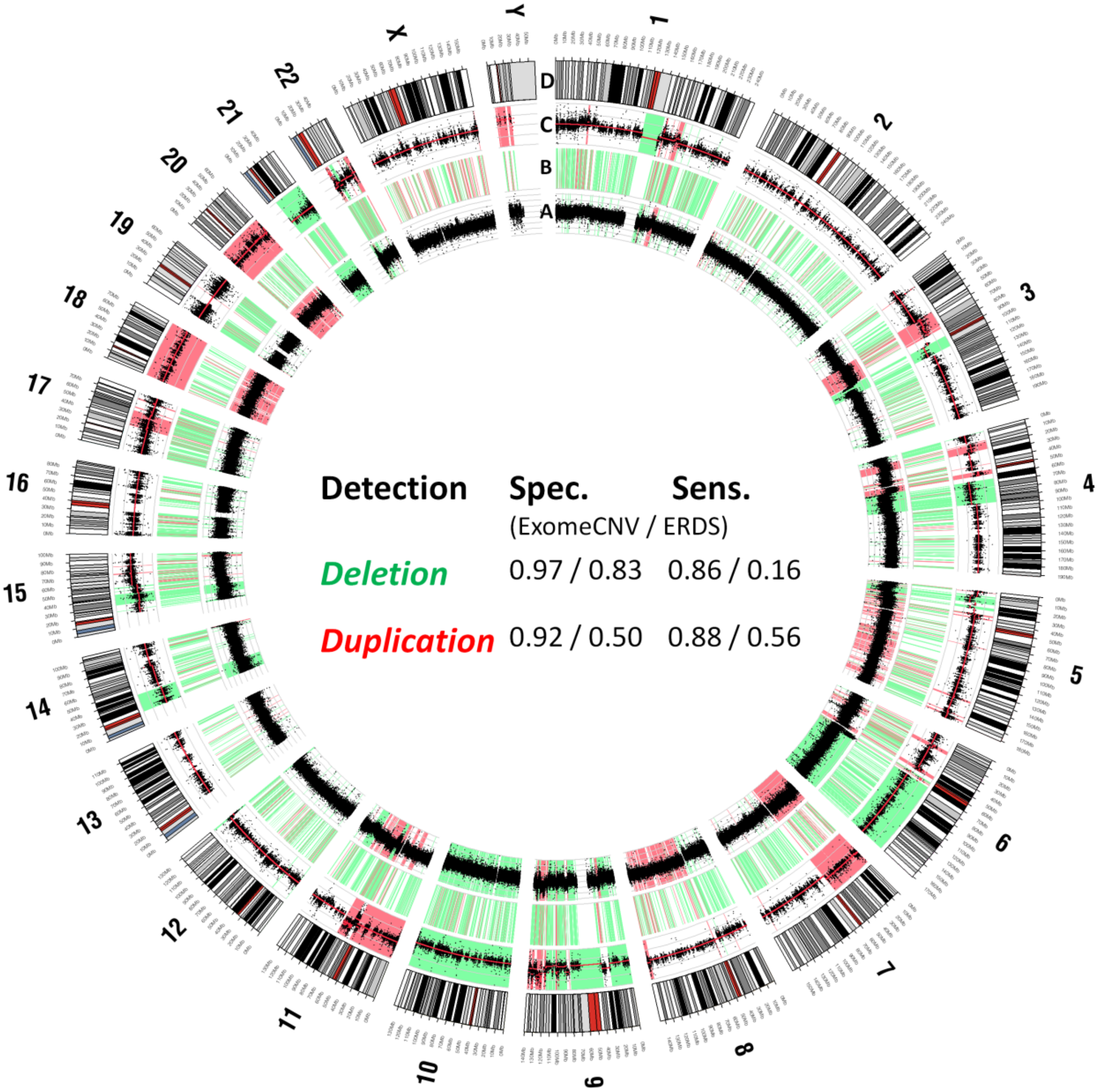


Figure S11 Sex-Chromosome Test for False Negative. Exons on Chromosome X are called as amplified by ExomeCNV with most (>99.9%) of the log depth-of-coverage ratio greater than 0. Only Chromosome X is shown here because almost all of the exons in Chromosome Y have depth-of-coverage of zero in male exome and cannot be meaningfully represented (the few exons with non-zero depth-of-coverage have very low depth-of-coverage, are called as deleted by ExomeCNV, and are sequencing errors). Different shades of red represent strength of the signal.

Since we knew the exact copy-number of sex-chromosomes in normal males and females, we used two internally available exomes, one male and one female, to test if ExomeCNV can detect this copy-number difference. The two individual exomes are from individuals with no evidence of sex-liked copy-number aberration. Treating the male exome as case and female as control, ExomeCNV correctly identified Chromosome X as being “amplified” and Chromosome Y as being “deleted” with no false negative (Figure S11). Here we set minimum specificity to 0.9999 in exon-wise calling and 0.9 in segment-wise calling (which is the default settings) and set admixture rate to 0.

Comparison between ExomeCNV and ERDS



Detection	Spec. (ExomeCNV / ERDS)	Sens.
Deletion	0.97 / 0.83	0.86 / 0.16
Duplication	0.92 / 0.50	0.88 / 0.56

Figure S12 Analysis of Melanoma and Paired Normal Samples. Comparison of CNV calls from exome sequencing data using ExomeCNV and ERDS, compared to calls from genotyping array. The most outer ring (D) shows the chromosome ideograms in a pter-qter orientation, clockwise with the centromeres in red. From inside to outside, each data track represents (A) Log R Ratio (LRR) from genotyping array with the region of gain highlighted in red and the region of loss highlighted in green; (B) CNV calls from ERDS, the region of gain highlighted in red and the region of loss highlighted in green (C) log ratio of tumor and normal depth-of-coverage with the segment mean in red line, the region of gain highlighted in red and the region of loss highlighted in green. The CNV for the chromosome Y were not called for the genotyping data as genoCN (the algorithm used to call CNV from Omni-1) is not designed to analyze chromosome Y. The table in the middle summarizes best achievable specificity and sensitivity of ExomeCNV and ERDS in detecting CNV relative to CNV calls from Omni-1 array assessment.

Using Pooled Sample as Control

In many applications, for example in identifying germline CNVs, we do not have matched normal sample to compare the exome of interest with. We propose using a pooled sample as control sample. Samples to be pooled have to be from libraries of the same type (e.g. paired-end, single-end) and processed by the same exon capture and sequencing protocols. This usually means all of the samples should be processed at a particular site. Pooling can be done by averaging depth-of-coverage of each exon across all exomes weighting each exome equally or by their total depth-of-coverage.

We have pooled eight available exomes which were processed and sequenced in the same manner as the melanoma samples used in our study. The depth of coverage on each exon was averaged and used as a “normal” control to compare against normal and tumor exomes from a melanoma patient. The variance of the pooled depth-of-coverage decreased, as expected by the central limit theorem (Figure S13). The ExomeCNV results are shown in Figure S14.

As cautioned in the main text, there is a problem in validating such analysis. First, we cannot ascertain that the pooled exome actually represents an exome with normal copy number of two. These exomes we used were from patients with various genetic abnormalities which may have abnormal copy number variants. Moreover, 4 of the 8 exomes came from the related individuals and may share substantial amount of germline CNVs which would skew the distribution of the pooled copy number. And even if all samples are from unrelated individuals, there might be common CNVs in the population that distorts the “normality” of the pooled sample.

Moreover, a CNV call from this analysis could arise because of many factors, and there is no direct way to validate the results. A CNV, say deletion, found by comparing the tumor exome against pooled exome may arise from 1) somatic deletion in tumor exome 2) germline deletion 3) duplication in pooled exome or 4) false positive. Since we do not have a germline CNV profiling of the subjects, there is no simple way to assess validity of this approach. Thus, we left this as a speculation in our discussion.

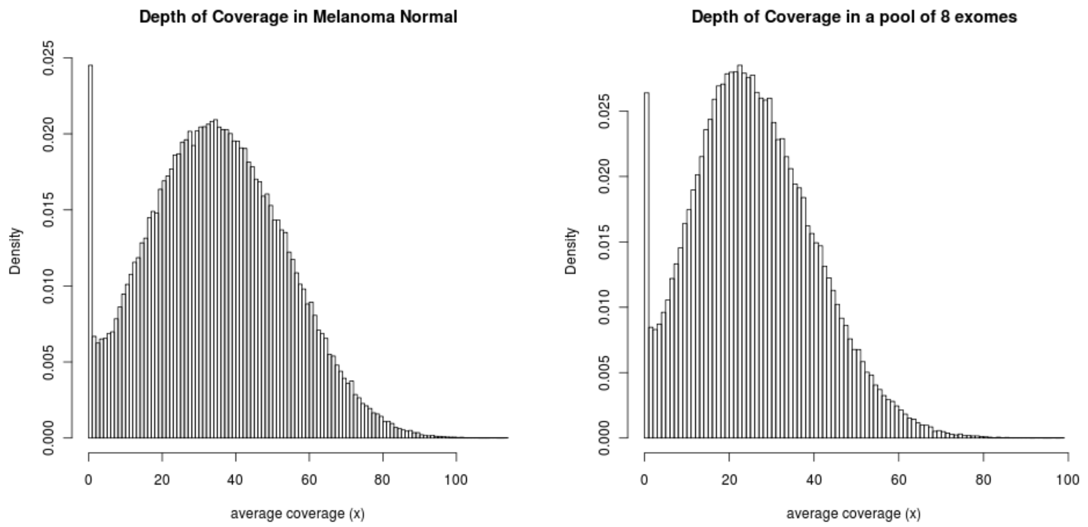
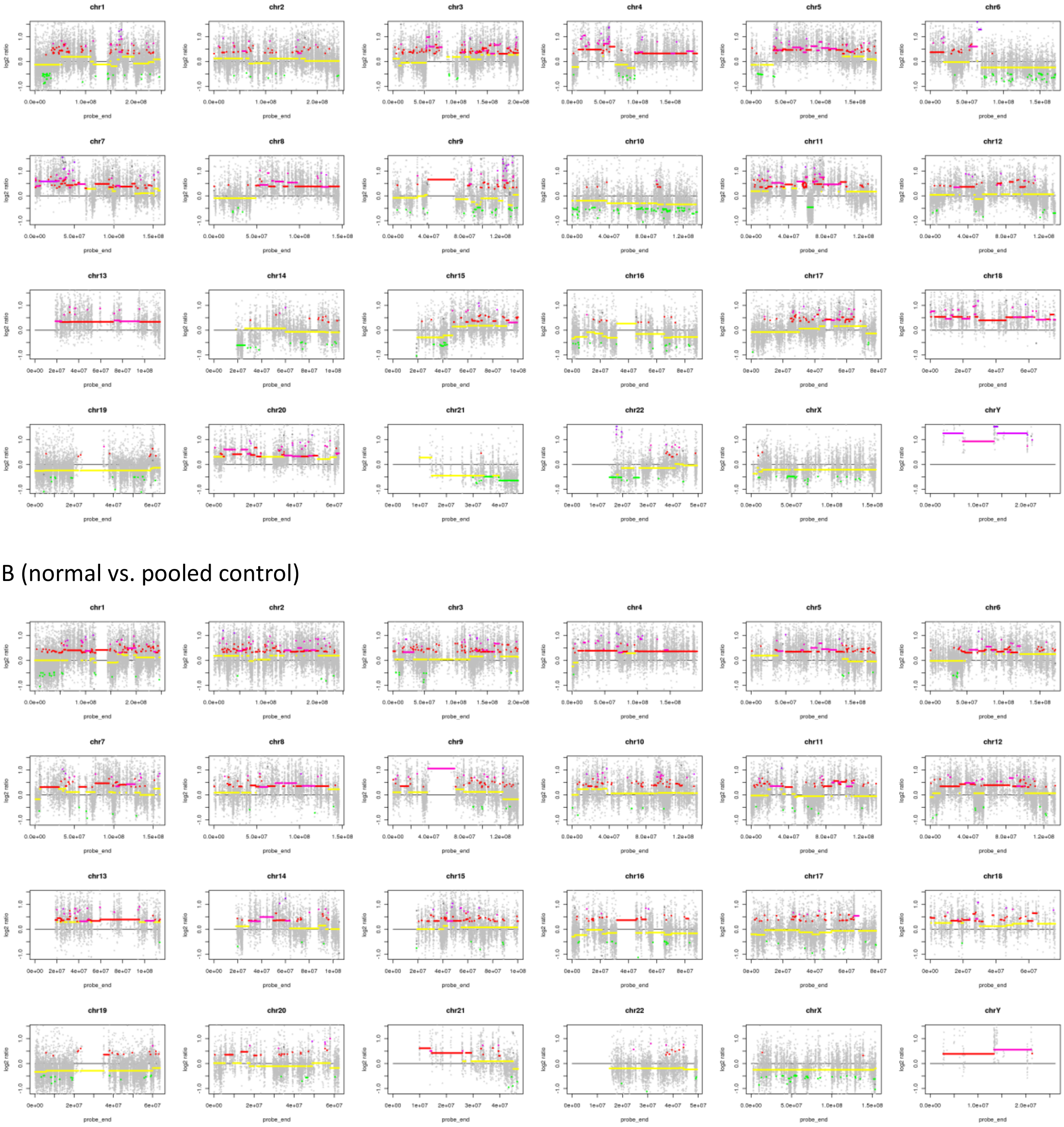


Figure S13 Pooling depth-of-coverage of as few as 8 exomes reduces the variance significantly.



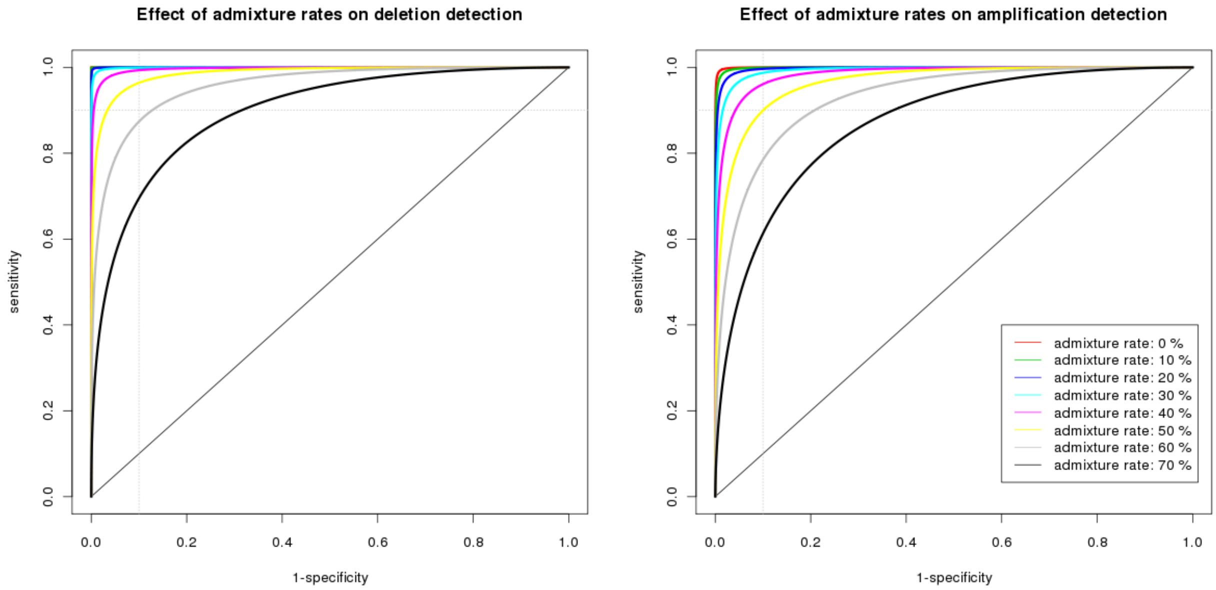
A (tumor vs. pooled control)

B (normal vs. pooled control)

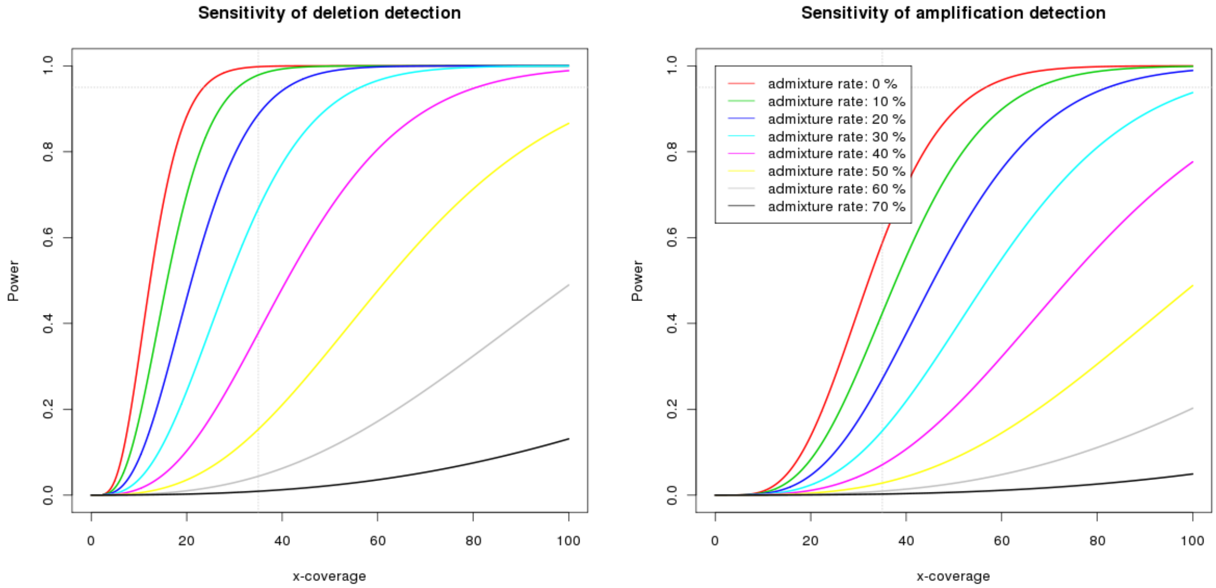
Figure S14 ExomeCNV results from comparing melanoma tumor (A) and matched normal (B) with the pooled sample. The amplification observed in chromosome Y is likely due to the imbalance composition of male/female in the pooled sample.

Effect of Admixture Rate on Sensitivity and Specificity of CNV Detection

A



B



C

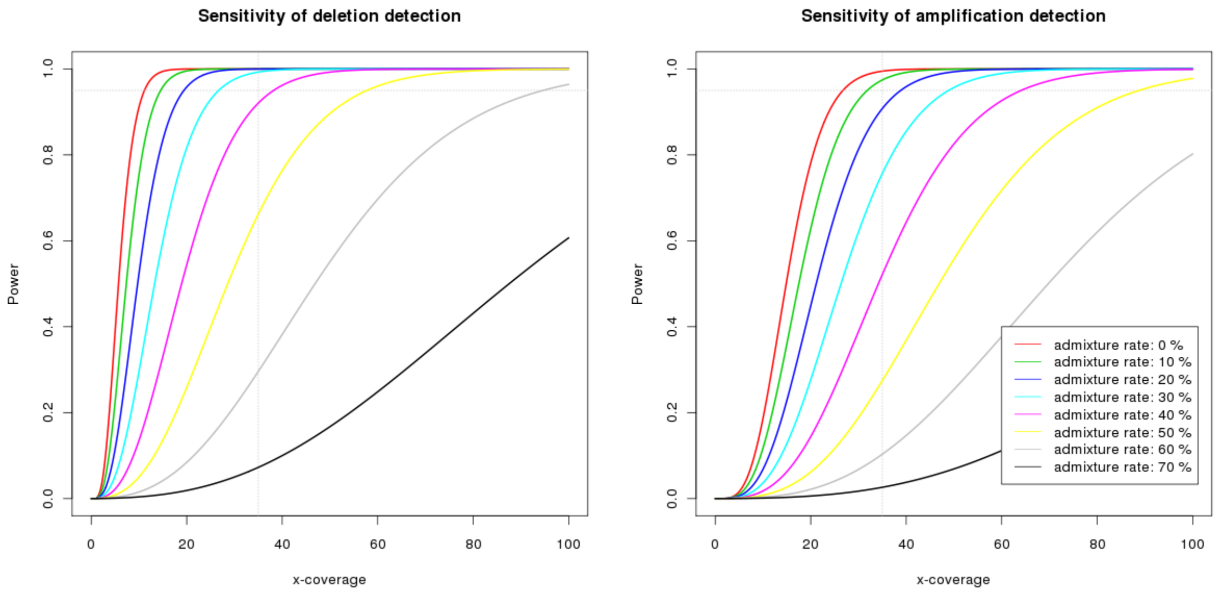


Figure S15 ROC and Power curves showing effect of varying admixture rates. (A) ROC curves showing sensitivity and specificity of detecting deletion and duplication of segments size 500bp. (B) Power curves for detecting CNV segments of size 500bp. Power is plotted relative to mean depth-of-coverage in the genomic segment, setting false positive to 1 per genome based on an analytical model of genome-wide power of detection. (C) Same as (B) but for detecting CNV segments of size 1000bp.

Estimation of Admixture Rate from LOH Regions

Because LOH detection method does not require prior knowledge of admixture rate, we can use LOH detection to estimate admixture rate. In particular, because we know that B-allele frequency (BAF) in a LOH region is either 0.5*c* or 1 - 0.5*c* where *c* is the admixture rate, the value $0.5 - |BAF - 0.5| = 0.5\ c$. Thus, *c* can be estimated by

$$\hat{c} = 1 - 2\,\mathrm{Average}(|BAF_{\mathrm{LOH}} - 0.5|)$$

where BAF_{LOH} is the B-allele frequency of a LOH region. In our melanoma sample, the admixture rate is estimated to be about 30% by this method, which is in agreement with an estimate from the SNP genotyping arrays.

Reference

1. Campbell, P.J., *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729 (2008).
2. Chiang, D.Y., *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, 99-103 (2009).